What if we took within-person performance variability seriously?

Fisher, Cynthia D.

Published in:
Industrial and Organizational Psychology

DOI:
10.1111/j.1754-9434.2008.00036.x

Licence:
Other

6-1-2008

What if we took within-person performance variability seriously?

Cynthia D. Fisher
Bond University, cynthia_fisher@bond.edu.au

What If We Took Within-person Performance Variability Seriously?

Cynthia D. Fisher

School of Business

Bond University

Gold Coast QLD 4229

Australia

Cynthia_Fisher@Bond.edu.au

Phone +61 7 5595 2215

Fax +61 5595 1160

What If We Took Within-person Performance Variability Seriously?

Efforts to understand what seems to be an unacceptably weak relationship between actual performance and rated performance have focused exclusively on the rater side of the model, not on the performance side. For instance, the Murphy model (in press) shows error only for ratings. Therefore, efforts to remedy the situation have also focused exclusively on raters: adjust the relationships of poor quality ratings to other variables for attenuation due to unreliability; improve the raters by training; clarify the rating task by providing a better format; or enhance rater motivation to be honest in recording what they really think. A strong implicit assumption underlies all of these approaches: that an employee's job performance is stable and that there is some true level of performance available to be observed and rated, if raters were just capable or motivated to do so. But what if part of the problem is that performance is not entirely stable over the short term? I will first establish that this is the case, then draw out some implications of true performance fluctuation for the relationship between performance and performance ratings.

Performance appraisal researchers readily accept some forms of instability, inconsistency, or intra-individual variation in true performance. For instance, most are happy to assume stable intra-individual variation across performance dimensions, hence the obsession with halo "error" in the older performance appraisal literature. There is also acceptance of longer term systematic changes in performance, reflected in the literature on the "dynamic criterion" and on efforts to model growth curves as new employees learn their jobs and improve over time. Finally, there is a sizable and growing literature on typical versus maximum performance, suggesting that some situations elicit better performance than others from the same individuals.

Appraisal researchers may be less aware of evidence that employee performance fluctuates within-person over short periods of time. A series of studies by Rothe (summarized in Rambo,

Chomiak, & Price, 1978) found that correlations between week to week output of factory workers in a variety of jobs varied from .48 to .82. Stewart and Nandkeolyar (2006) observed the weekly sales performance of 167 salespeople for 26 weeks, and found that 73% of the variance in weekly sales was within-person. Fisher and Noble (2004) prompted 121 employees to report their task performance five times per day for two weeks, and found that 77% of the variance in self-ratings of momentary performance was within-person. Clearly, individuals do not perform, or see themselves as performing, at exactly the same level at all times.

This shouldn't be a surprise. Attribution theory suggests unstable causes of performance, such as effort, luck, and fatigue. Other influences on performance that are likely to fluctuate over time in most work settings include task complexity, task interdependence, task priority, and environmental opportunities or constraints. This is not to deny the existence of very stable (intelligence, personality) and fairly stable (skill, job knowledge) causes of average performance, but rather to point out that there are also genuine shorter term causes of performance fluctuation.

Some rating formats acknowledge variation in performance by requiring raters to report the frequencies with which employees exhibit specific behaviors. The performance distribution assessment method (Kane, 1986) goes farther by asking raters to report the percent of time that a ratee displays each of five levels of performance, from minimum possible to maximum possible, on each job function. Borman (1991) discusses intra-individual variability in performance in a chapter on performance criteria. He states (p. 277), "it seems obvious that employees' performance on individual dimensions varies over time and across different situations on the job…performance is probably more faithfully characterized by a distribution than it is by a single number. " He concludes (p. 279), "performance distributions may provide comparatively rich

descriptions of individuals' performance, giving us substantially greater understanding of that performance, along with its causes and consequences."

Despite these warnings, the notion that performance genuinely fluctuates over short periods of time has not sunk deep into the awareness of most appraisal researchers, and is not often considered by those who study rater error. Short term within-person variation tends to be regarded as error, when in fact it may have substantive causes and be predictable by other transient features of the employee or the work environment (c.f. Fisher & Noble, 2004). Below, I will explore a few of the implications of performance fluctuations for the three models of the performance-performance rating relationship discussed by Murphy.

One-Factor Models: It's all Rater Error

If performance is defined as the total contribution made by a person over a year of work, then perhaps it is reasonable to regard fluctuations around some mean level of contribution by an individual as error. However, in discussing transient error in ratings, Schmidt and Hunter (1990) lay the blame entirely on the rater, e.g. on mood that temporarily biases a rater's responses. This ignores the possibility that poor test-retest reliability may not be rater error at all, but an accurate perception of a phenomenon that is less than perfectly stable. Inter-rater agreement may be weak because raters have legitimately observed different episodes of performance at different levels. In fact, Woehr and Miller (1997) have demonstrated that interrater agreement is lower when ratee performance varies more within dimension.

To draw a parallel with intelligence testing, assume for the moment that the intelligence of an individual did truly vary over the short term, albeit around that person's characteristic mean level. If this were the case, we would not blame the intelligence test for being unreliable if it returned somewhat different readings at different points in time. Instead, we would accept that

the phenomenon being measured was varying and that the instrument may have correctly captured that variation. If performance truly does fluctuate over time and raters notice and capture current performance in their ratings, then estimates that rater error accounts for about half the variance in performance ratings (Viswesvaran, Ones, & Schmidt, 1996) may be inflated. Maybe raters are doing a better job than we think they are. If so, then correcting relationships between ratings and other variables for attenuation due to substantial rater unreliability would overestimate these relationships.

Multi-Factor Models: Cognitive Limitations of Raters and Other Non-Performance Factors Systematically Contaminate Ratings

Motowidlo, Borman, and Schmit (1997, p. 72) define performance as, "the aggregate value to the organization of the discrete behavioral episodes that an individual performs over a standard interval of time." This definition highlights the occurrence of performance in episodes, and as previously established, these episodes probably fluctuate in quality over time. Most rating tasks require the retrospective assessment of behavior over a full year, probably on the basis of dimensions which are not naturally used to encode performance when observing episodes in real time. The rating task does not correspond to the raw event-level data available to observers, nor to the way that event-level data are organized in memory. It should not be surprising that raters cope poorly with these demands and are subject to a number of biases in consequence.

We know that individuals have cognitive limitations which cause them to forget many discrete events. The research on autobiographical memory documents poor recall of repeated "mundane" events. Memory for someone else's everyday events should be even worse. There is also evidence that events are not equally weighted in final overall ratings. Primacy effects (perhaps due to cognitive categorization and a search for confirming evidence) and recency

effects (due to availability) are examples of unequal weighting. Research on retrospective reporting of affect and pain over time suggests that "peak" and "end" experiences often contribute to retrospective ratings over and above the mean of ratings collected in real time. Given some fluctuation in performance over time, it seems likely that raters will similarly be affected by the worst and/or the best performance episodes in the time period in question. Newman, Krzystofiak, and Cardy (1986) found that greater intra-dimensional variance in performance (but within the same either high or low end of the performance dimension) was associated with higher performance ratings. This could be interpreted as a "peak" effect, where the best performance episode has a disproportionate influence on the overall rating.

However, are these effects error? The assumption that true performance should be conceptualized as the mean of all equally weighted performance episodes may be incorrect. Raters who diverge from this mean by incorporating an aspect of variability into their ratings (such as best or worst performance) may be reporting more accurately on a ratee's actual worth to the organization. For instance, Kane (1986) discussed the importance of "negative-range avoidance." Two employees of equal average performance may not be of equal value to the organization if one occasionally commits a truly serious error. Extent of variability itself may be an important property of performance. As an example, Barnes and Morgeson (2007) have shown that the value of professional basketball players to their employers (operationalized as following year salary) is negatively predicted by shooting variability across games. Others have recently suggested that stable individual differences may predict which employees are able to maintain performance consistency despite fluctuating environmental conditions (Mangos, Steele-Johnson, LaHuis, & White; Stewart & Nandkeolyar, 2007). Perhaps organizations should acknowledge and assess both mean performance and performance variability.

Mediated Models: Rater Motivation Creates Intentional Distortion

Murphy's third model suggests that raters create error by intentionally rating individuals differently than their true performance, for the purpose of sending a message, motivating ratees, maintaining relationships, or managing impressions. The underlying assumption is that raters do know how well ratees are performing, but need to be motivated to record it accurately. It probably is true that rater motivation accounts for some intentional distortion in ratings. However, given variability in a ratee's performance over time, it is also likely that a rater can legitimately point to discrete performance episodes that are consistent with any rating that he or she chooses to give. If the motivation to give a particular rating comes first, then the rater will be particularly vigilant for the subsequent occurrence of performance episodes, however rare, that corroborate the desired rating. Raters may genuinely believe the distorted ratings they give, based on the performance episodes they selectively observe and recall.

Conclusions and Further Implications

Murphy wondered whether there had been "a faulty diagnosis of the problems that beset performance appraisal." I argue that the diagnosis has been somewhat deficient in focusing entirely on the rater while failing to consider that performance itself is a moving target. Performance variation over short periods of time does exist within-person and within-dimension. Our definitions of the construct of job performance should more explicitly acknowledge this fact. The construct definition will then drive the conceptualization of error in ratings. The definition might also spark a search for different performance measurement approaches.

We know that the less stable a phenomenon, the more samples or observations are required to accurately estimate a mean. Perhaps one measurement solution would be to rate global short term performance at the episode level, then aggregate (either mechanically or judgmentally) over

the year. The classic idea of raters keeping diaries is consistent with this suggestion of measuring performance episodes as they occur. Raters would be asked to assess how good a particular performance episode was overall, rather than being asked to perform the difficult task of decomposing an observed performance episode into performance dimensions. Further, memory would be much less of an issue as raters are not asked to recall and integrate across many and varied incidents over long time periods. Transient error due to rater mood would be washed out across multiple observations. The motivation to distort ratings would be reduced, as the evaluation of any single episode has relatively minor impact on the final composite. It would be interesting to find out whether interrater agreement increased under such a rating scheme.

Recording assessments of all performance episodes would be a very time consuming task, so an alternative would be to create a representative sampling plan on when to elicit ratings of episodes. Collection of this type of repeated measurements over time would allow for the calculation of other metrics in addition to mean performance. As suggested above, best, worst, and performance variability could also be assessed, and may be useful data for an organization to have in considering the true worth of an individual's contribution over time. The labor intensive nature of this approach suggests that it might be useful only when highly accurate performance measures, or performance variability measures, are needed. One such setting might be selection test validation. Undoubtedly there are many more implications of true short term within-person performance variance for performance appraisal, which I leave for other scholars to draw out in the future.

References

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance, 20*, 259-274.

Borman, W. C., (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough, (Eds). *Handbook of industrial and organizational psychology*, *Vol. 2 (2nd ed.)* (pp. 271-326). Palo Alto, CA: Consulting Psychologists Press.

Fisher, C. D., & Noble, C. S. (2004). A within-person examination of correlates of performance and emotions while working. *Human Performance*, 17, 145-168.

Hunter, J.E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Kane, J. S. (1986). Performance distribution assessment. In R.A. Berk, (Ed). *Performance assessment: Methods & applications* (pp. 237-273). Baltimore, MD: Johns Hopkins University Press.

Mangos, P. M., Steele-Johnson, D., LaHuis, D., & White III, E.D. (2007). A multiple-task measurement framework for assessing maximum-typical performance. *Human Performance*, *20*, 241-258.

Motowidlo, S. J., Borman, W. C, & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71-83.

Murphy, K.R. (in press). Explaining the weak relationship between job performance and ratings of job performance.

Newman, J., Krzystofiak, F., & Cardy, R. (1986). Role of behavior level, behavioural variability, and ratee order in the formation of appraisal ratings. *Basic and Applied Social Psychology, 7*, 277-293.

Rambo, W.W., Chomiak, A.M., & Price, J.M. (1978). Consistency of performance under stable conditions of work. *Journal of Applied Psychology, 68*, 78-87.

Stewart, G. L., & Nandkeolyar, A. K. (2006). Adaptation and intraindividual variation in sales outcomes: Exploring the interactive effects of personality and environmental opportunity. *Personnel Psychology, 59*, 307-332.

Stewart, G. L., & Nandkeolyar, A. K. (2007). Exploring how constraints created by other people influence intraindividual variation in objective performance measures. *Journal of Applied Psychology, 92*, 1149-1158.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Woehr, D.J., & Miller, M.J. (1997). Distributional ratings of performance: More evidence for a new rating format. *Journal of Management, 23*, 705-720.